2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 978-1-6654-8402-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/VRW55335.2022.00194

How Late is Too Late? Effects of Network Latency on Audio-Visual Perception During AR Remote Musical Collaboration

Torin Hopkins* Suibi Che-Chuan Weng† Rishi Vanukuru‡ Emma Wenzel§ Amy Banic** ¶
Ellen Yi-Luen Do ‖

ATLAS Institute, University of Colorado Boulder
**Interactive Realities Lab, University of Wyoming

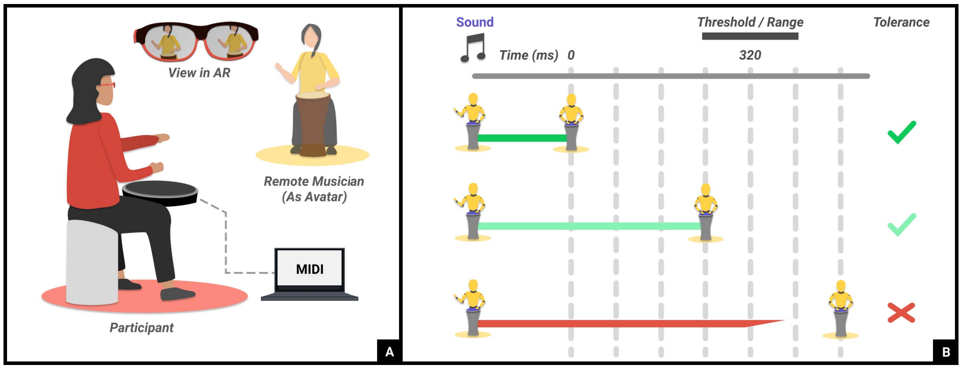

Figure 1: A) A representation of the experimental setup involving a participant collaborating with a remote musician using AR glasses and a MIDI drum, and B) An illustration of the sound-animation latency threshold being studied.

Abstract

Networked Musical Collaboration requires near-instantaneous network transmission for successful real-time collaboration. We studied the way changes in network latency affect participants' auditory and visual perception in latency detection, as well as latency tolerance in AR. Twenty-four participants were asked to play a hand drum with a prerecorded remote musician rendered as an avatar in AR at different levels of audio-visual latency. We analyzed the subjective responses of the participants from each session. Results suggest a minimum noticeable delay value between 160 milliseconds (ms) and 320 ms, as well as no upper limit to audio-visual delay tolerance.

Index Terms: Human-centered computing—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Interaction paradigms—Collaborative Interaction; Applied computing—Sound and music computing

1 Introduction and Related Works

Networked Extended Reality (XR) collaborative systems inevitably encounter the issue of network latency—defined as the time it takes data to transfer from its original source and its destination. Networked XR experiences often require reconciliation of several types of information (audio, visual, haptic, etc.) to be perceived as a unified event [1, 3]. Networked musical experiences are examples of activities that require very low latency to function [2, 9]. For short distances, networked musical experiences using audio-only platforms can achieve the low-latency required for music collaboration [4, 6]. Visuals, however, take a longer time to traverse a network because they require more processing and contain more data than in audio-only systems. This results in a delay between audio and visual information, which can make musical collaboration difficult due to the cue mismatch. Attempts to overcome mismatching cues have been explored using many forms of Extended Reality (XR) [5, 7, 8, 10], but none to date have accomplished real-time collaboration in a musical context. Using AR, we aim to gauge the minimum noticeable delay perception between visual and auditory stimulus, as well as the tolerability of the delay experienced. This enables us to design a networked XR musical experience with a known tolerance for latency between audio and visuals.

2 Method

To answer our research questions, we conducted a 2x8 within-subjects experiment to simulate remote musical collaboration in AR. We experimentally controlled for two types of remote drumming collaboration: Mimic and Improvise, and eight levels of network latency: 0ms, 20ms, 40ms, 80ms, 160ms, 320ms, 640ms, and 1200ms. We used Balanced Latin Squares across all conditions to reduce potential ordering effects. We first discuss the tasks and apparatus, followed by details about the participants involved and the experimental procedure. In both cases, the speed of the rhythm played by the remote partner was 90 beats per minute. During the "mimicry task", the remote partner would play a strict 4/4 (4 beats per bar, 4 bars per measure) rhythm, which translates to a gap of 750 ms between subsequent hits of the drum. In the "improvise" task, the remote player would continue playing at 90 beats per minute, but using a more natural, free-form rhythm. 24 Participants used the Nreal Light headset to view their remote partner in AR. Participants were given an electronic drum pad to play along with the avatar. An 18-camera Qualysis motion capture system[1] was used to record

*e-mail: torin.hopkins@colorado.edu
†e-mail: chwe1250@colorado.edu
‡e-mail: rishi.vanukuru@colorado.edu
§email: emma.wenzel@colorado.edu
¶email: amy.banic@colorado.edu
‖email: ellen.do@colorado.edu

[1] https://www.qualisys.com/

978-1-6654-8402-2/22/$31.00 ©2022 IEEE
DOI 10.1109/VRW55335.2022.00194

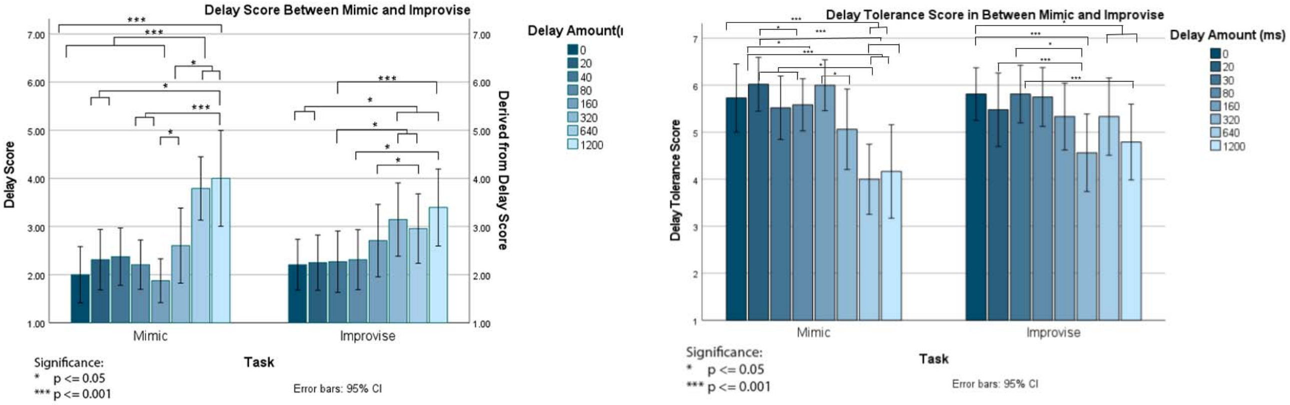

Figure 2: The bar chart on the left demonstrates the difference in subjective delay score between the mimic and improvise tasks. Similarly, the bar chart on the right indicates the difference in reported delay tolerance between the mimic and improvised tasks.

player movement as animation clips. The drum audio was recorded into the Ableton digital audio workstation software via an external sound card. Once recorded, these motion and audio files were applied to a Humanoid avatar (sourced from the Mixamo character and animation platform[2]) and rendered via an application for the Nreal Light AR headset [3] using the Unity live development engine (version 2020.2.2f1). This application was presented to participants, with differing offsets between the music and animation (Figure 1).

3 Results and Discussion

Our results were derived by assessing a quantitative score between 1 and 7 for two questions (How much delay did you experience between animation and sound?, and How would you rate the tolerability of the delay experienced?). To analyze the data, a repeated measures Analysis of Variance (RM-ANOVA) was conducted on the data. Mauchly's test did not reveal a violation of sphericity for any of the RM-ANOVA tests. Post-hoc Least Significant Difference (LSD) tests explained any significance among the groups.

In this study we aim to define a threshold for both perceiving a minimum noticeable delay and when players can no longer tolerate the amount of delay between a perceived sound and visual animation. We addressed the first research question of defining the minimum noticeable delay by assessing the delay amount scores indicated by the participants after each task. Results indicated that players began significantly noticing delay between 160ms and 320ms (A RM-ANOVA revealed a significant difference among minimum noticeable delay means, F(7, 161)= 9.74, p<0.001, η_p^2=0.30, where the delay amount conditions of 0ms, 20ms, 40ms, 80ms, 160ms, were significantly smaller than delay amounts of 320ms, 640ms, 1200ms (see Figure 2, left). Therefore, we hypothesize that the minimum noticeable delay value is higher than 160ms and possibly less than 320ms. In future studies we will investigate perceived delay between 160ms and 320ms to test for a more precise value of minimum noticeable delay and control for musical experience.

We address the maximum amount of delay a participant can tolerate while playing drums. We noticed that the tolerance began significantly declining at the same delay amount of 320ms as the reported delay score. This suggests that as soon as players begin noticing delay, the experience begins to degrade. Interestingly, the reported delay tolerance levels are not observed below a neutral score of 4 (participants began to score the tolerance lower at 320ms of delay F(7, 161)= 9.94, p<.001, η_p^2=.30, where the delay amount conditions of 0ms, 20ms, 40ms, 80ms, 160ms, were significantly smaller than delay amounts of 320ms, 640ms, 1200ms (see Figure 2, right). We interpret this to mean that a maximum threshold for delay was not found during our test sequence. Because players were mimicking a virtual drummer playing at a regular tempo of 90bpm (750ms between beats), 1200ms of delay causes animations to fall out of sync by almost a full cycle of sequential beats. This poses a limit for understanding an upper threshold for delay tolerance with regularly spaced drum beats because a player will not be able to distinguish the discrepancy in delay between cycles.

In summary, for designing collaborative music augmented reality applications, visual adjustments should be made during the minimum noticeable delay range between 160 milliseconds (ms) and 320 ms. Further studies, including those that investigate irregular drum beats, are needed to understand audio-visual delay tolerance upper limits.

References

[1] W. J. Adams. The development of audio-visual integration for temporal judgements. *PLOS Computational Biology*, 12(4):e1004865, 2016.

[2] E. Chew, A. Sawchuk, C. Tanoue, and R. Zimmermann. Segmental tempo analysis of performances in user-centered experiments in the distributed immersive performance project. In *Proceedings of the Sound and Music Computing Conference, Salerno, Italy*, 2005.

[3] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor. Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, 36(38):9888–9895, 2016.

[4] J.-P. Cáceres and C. Chafe. Jacktrip: Under the hood of an engine for network audio. *Journal of New Music Research*, 39:1–4, 09 2010. doi: 10.1080/09298215.2010.481361

[5] Z. He, R. Du, and K. Perlin. Collabovr: A reconfigurable framework for creative collaboration in virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 542–54.

[6] Jamulus. Jamulus – Play music online. With friends. For free.

[7] L. Men and N. Bryan-Kinns. Lemo: supporting collaborative music making in virtual reality. In *2018 IEEE 4th VR workshop on sonic interactions for virtual environments (SIVE)*, pp. 1–6. IEEE, 2018.

[8] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti. An overview on networked music performance technologies. *IEEE Access*, 4:8823–8843, 2016.

[9] N. Schuett. The effects of latency on ensemble performance. *Bachelor Thesis, CCRMA Department of Music, Stanford University*, 2002.

[10] L. Turchet, R. Hamilton, and A. Camci. Music in extended realities. *IEEE Access*, 9:15810–15832, 2021.

[2] https://www.mixamo.com/

[3] https://www.nreal.ai/light/